WHEN WAS IT WRITTEN?

Prediction of publication date based on article content.

CS 229 Final Project

Joseph Bakarji (jbakarji@stanford.edu), Dimitrios Belivanis (dbelivan@stanford.edu), Sepehr Nezami (nezami@stanford.edu)

INTRODUCTION

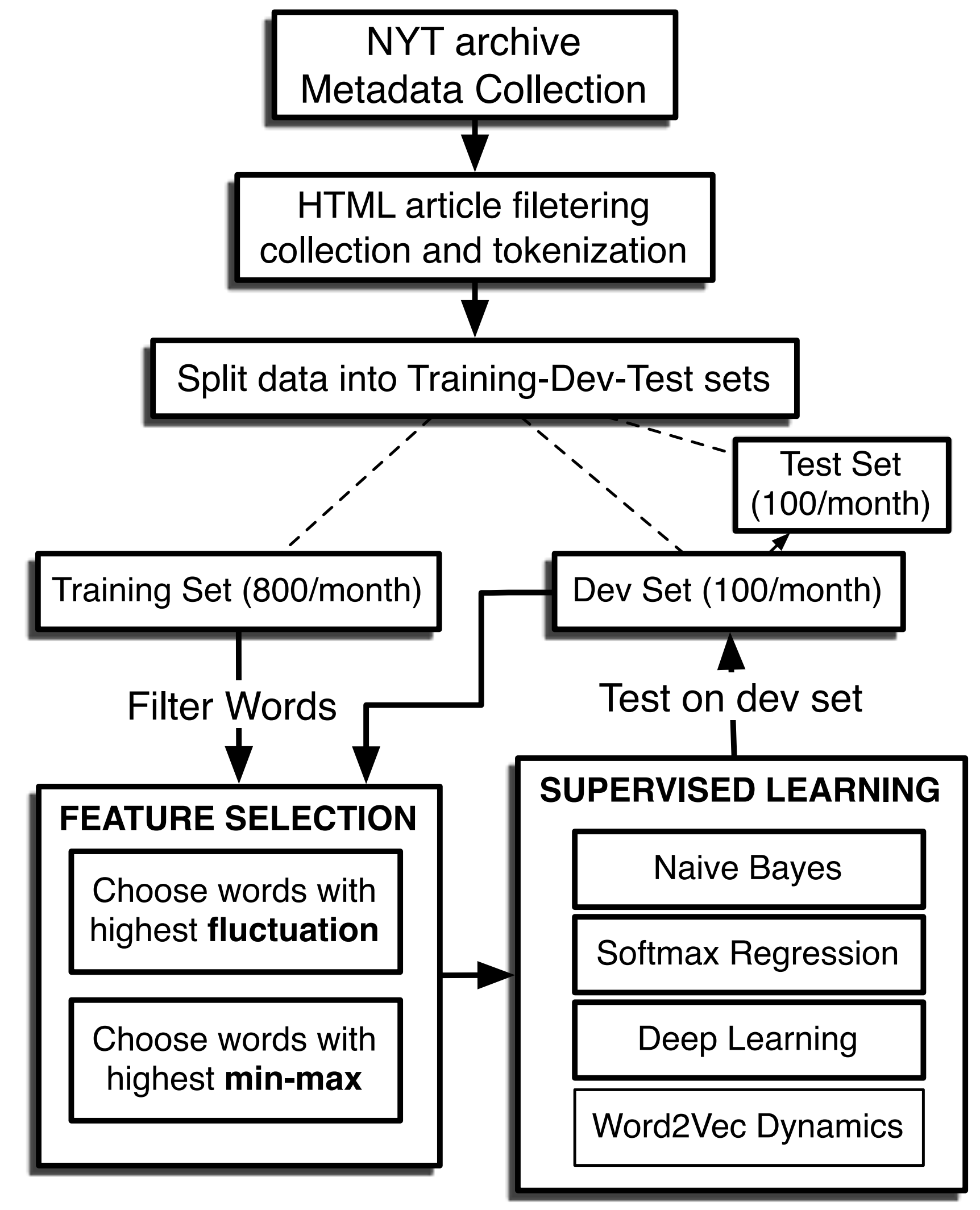
Can the language used in an article, blog post or book give us any clues about their publication date?

In this project, we develop a supervised learning algorithm that reveals and learns the correlation between the content of an article and its publication date. The result is used to date articles without date-stamp. E.g., assign a date to news spinets without reference, categorize articles according to their relevant time periods.

The motivation is the irregular use of certain words as seen in news articles and Google analytics.

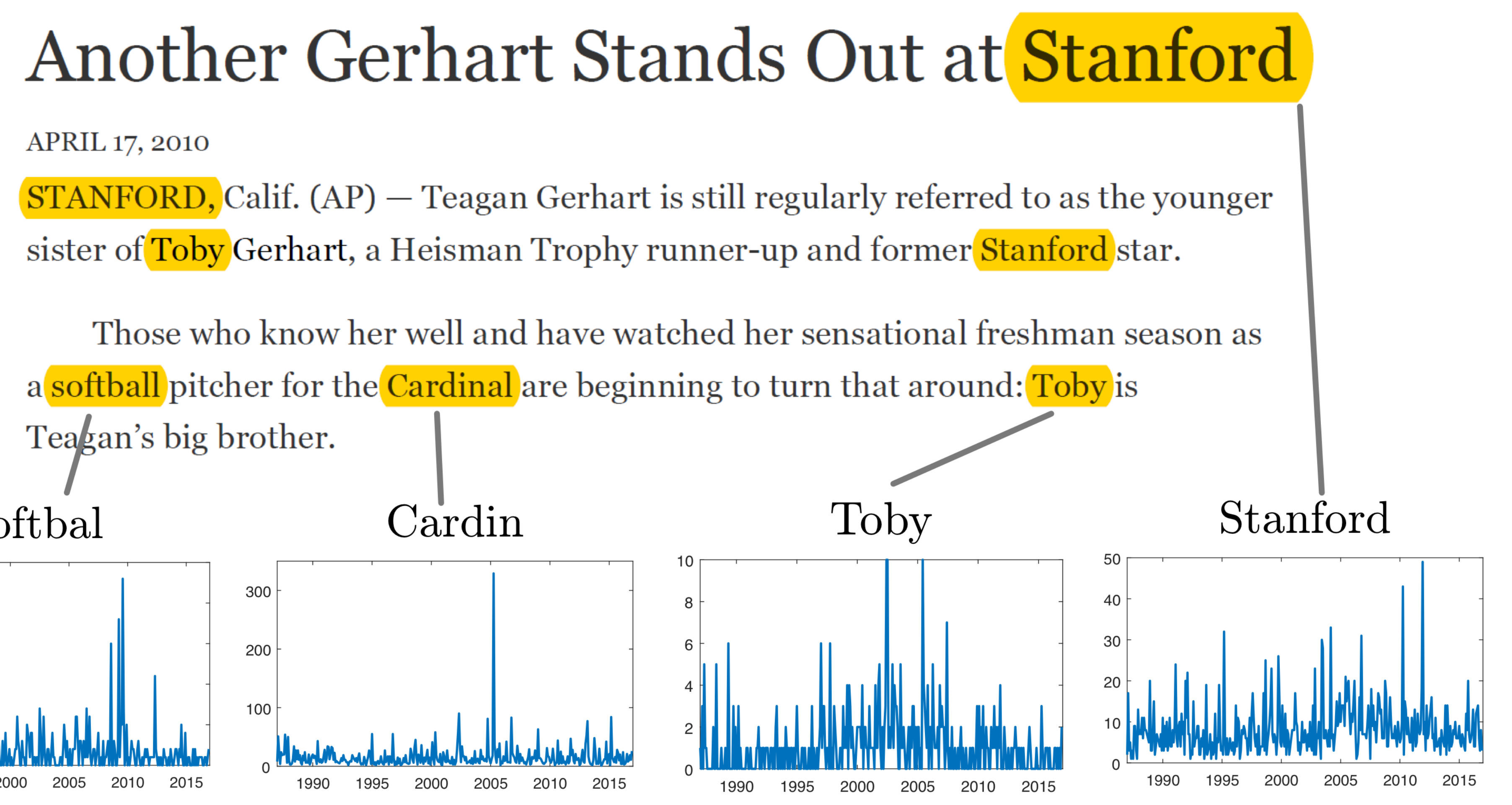
METHODOLOGY

Having access to the New York Times API, collected the content of a 1000 articles/month for 30 years (1987 to 2017). First we tokenize, stem and filter words that would not be helpful features. The final step in data collection was filtering out articles with too little content (such as video or slide-show posts).



The main step before learning is feature selection, i.e. choosing words whose frequency reveals specific events in time. We then iterate between feature selection and learning using SVM, Logistic Regression, Deep Learning, and Naive Bayes. We did the learning on both a monthly and yearly basis.

FEATURE SELECTION



Another Gerhart Stands Out at Stanford

APRIL 17, 2010

STANFORD, Calif. (AP) — Teagan Gerhart is still regularly referred to as the younger sister of Toby Gerhart, a Heisman Trophy runner-up and former Stanford star.

Those who know her well and have watched her sensational freshman season as a softball pitcher for the Cardinal are beginning to turn that around: Toby is Teagan's big brother.

Due to the large number of features (or words) a heuristic method for feature selection is used. A measure of fluctuation or average velocity of word appearance is given by:

$$\text{score}_v(\text{words}) = C_w \frac{\sum_{y=1987}^{2016} |f_w(y+1) - f_w(y)|}{\sum_{y=1987}^{2016} f_w(y)}.$$

$f_w(y)$ the frequency of the word w in year y, and proportionally constant C_w gives more weight to frequent words.

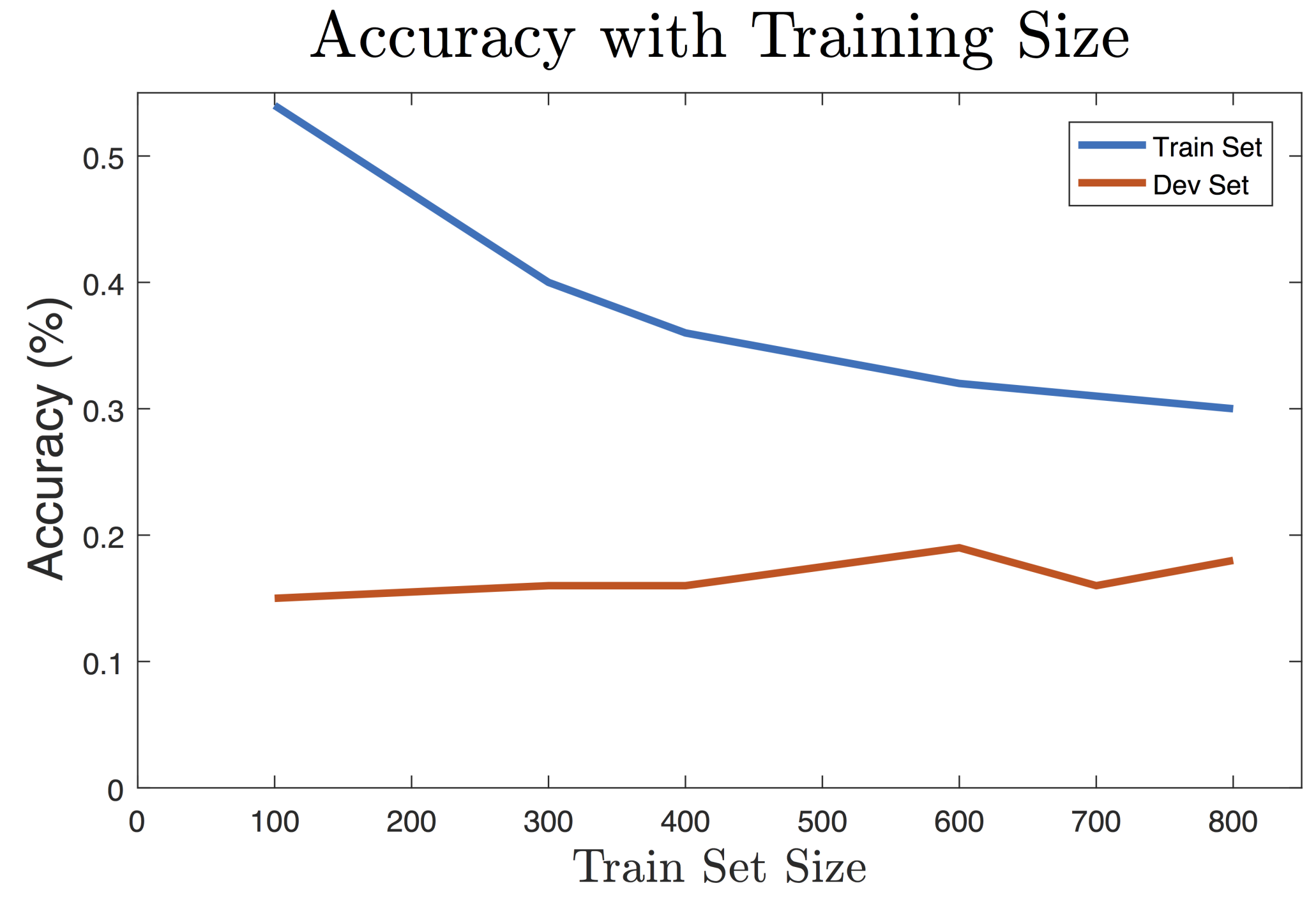
An iterative method for feature refinement is used based on the relative success of words to predict the given year, according to

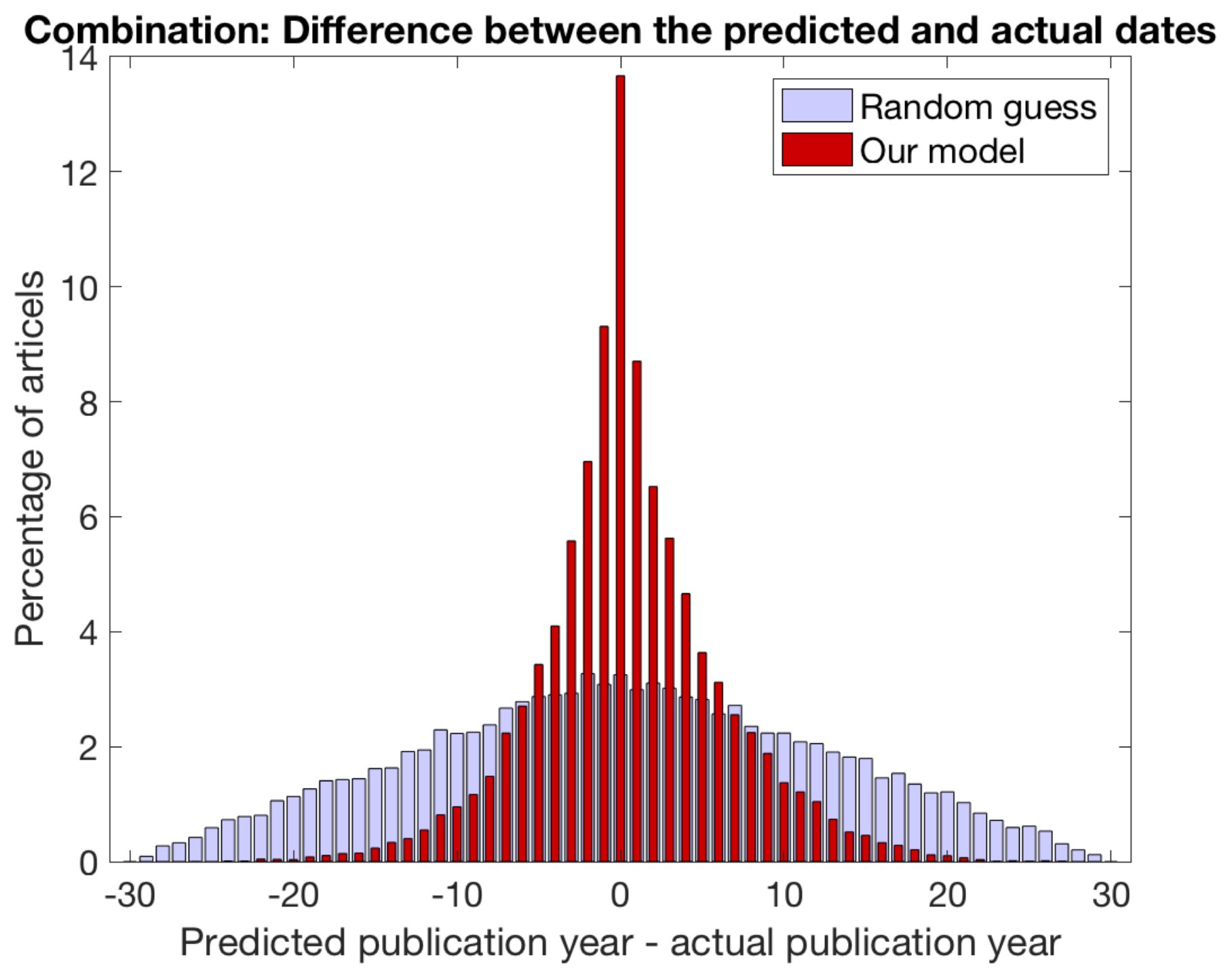
$$\text{score}_p(\text{word}) = \frac{P(\text{word}|\text{correct-pred})}{P(\text{word}|\text{false-pred}))}$$

Feature selection is a challenge when the corpus contains around 300K words. More features causes over fitting while less causes bias.

RESULTS

We study the error of our model as a function of size of the training set for Naive Bayes classifier. It can be seen that with the current size of training set, both 1-year prediction error and standard deviation of error have reached their asymptotic values.





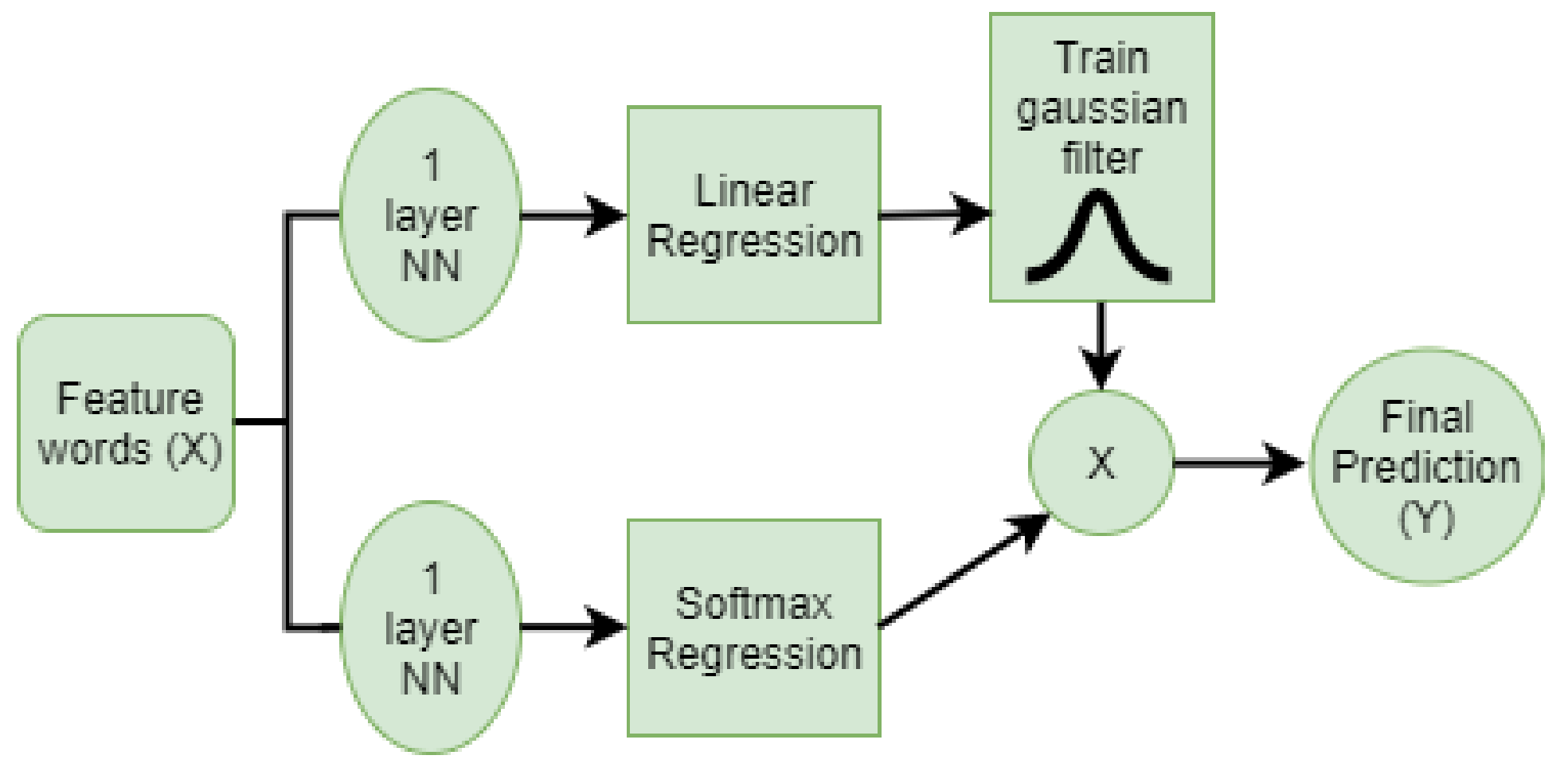
It can be seen that the predictability improves when linear regression and Naive Bayes are used in combination. Logistic regression doesn't do better on any of the sets due to the large number of features.

MODELS EVALUATION

Our preliminary results are given by the following table. Better ones are coming as we speak!

Model	Accuracy Training Set	Accuracy Dev. Set
NB-sklearn	35%	24%
NB-MAT	31%	29%
LinR-MAT	21%	9%
LR	40%	22%
NN-SM	22.4%	11.8%
NN-LinR	5.6%	5.5%

DEEP LEARNING



The accuracy of the NN with linear regression is low but provides information for the correct prediction vicinity. This information will be modeled as a Gaussian pdf and with the combination of Softmax prediction the accuracy is boosted.

CONCLUSION & FUTURE WORK

We showed that there is a correlation between the content of news articles and their publication date. Naive Bayes did better than NN and Softmax. Feature selection is the biggest challenge. In the future We would like to:

- Use mutual information and for feature selection
- Word2Vec and k-means for studying the dynamics of meaning and for reducing the number of features.
- Using n-grams and focusing more attention on names.
- Study how the error improves with the time range and across source.